Finding the Most Influential Student
A Random Walk Approach

Vrushank Ahire
2022CSB1002

April 25, 2024

Abstract

A method for identifying the most influential node, termed the "Super Winner", in a network of students is presented based on a random walk algorithm. The network is represented as a directed graph, where nodes correspond to students, and edges represent relationships between them. By simulating a random walk on the graph and analyzing visit frequencies of nodes, the Super Winner is determined as the student exerting the highest influence within the network. A mathematical explanation of the random walk process and transition probabilities is provided, along with an analysis of the algorithm's performance and potential applications in social network analysis and recommendation systems.

1 Introduction

The analysis of social networks has become increasingly important in various domains, including marketing, education, and community development. One of the key challenges in social network analysis is identifying the most influential individuals within a network, often referred to as the "Super Winner" or "Influencers". These individuals have a significant impact on the flow of information and the propagation of ideas within the network.

In this report, we focus on identifying the Super Winner in a network of students based on their relationships and interactions. The network data is represented as a directed graph, where each node corresponds to a student, and each edge represents a relationship between two students. Our approach utilizes a random walk algorithm to simulate the movement of a "walker" through the network, and the Super Winner is identified as the student (node) with the highest visit frequency during the random walk process.

2 Related Work

Random walk algorithms have been widely used in various fields, including computer science, physics, and social network analysis. In the context of social networks, random walks have been employed for tasks such as community detection [4], link prediction [3], and influence maximization [1].

Specifically, the concept of identifying influential nodes in a network has been explored in several studies. Kleinberg et al. [2] introduced the HITS algorithm for ranking web pages based on their authority and hub scores, which can be interpreted as a form of influence analysis. Kempe et al. [1] proposed a greedy algorithm for influence maximization in social networks, aiming to identify a set of initial nodes that can maximize the spread of information or influence.

3 Mathematical Analysis

Let $G = (V, E)$ be the directed graph representing the student network, with $n = |V|$ nodes. Let P be the $n \times n$ transition probability matrix of the random walk on G, where

$$p_{ij} = \begin{cases} \frac{1}{d_i}, & \text{if } (i, j) \in E \\ \frac{1}{n}, & \text{if } d_i = 0 \text{ and } i = j \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

and d_i is the out-degree of node i.

Let $\pi^{(t)}$ be the row vector representing the probability distribution of the random walker being at each node after t steps, with $\pi^{(0)}$ being the initial distribution satisfying $\sum_{i=1}^{n} \pi_i^{(0)} = 1$. Then,

$$\pi^{(t+1)} = \pi^{(t)} P \tag{2}$$

The stationary distribution π^* satisfies

$$\pi^* = \pi^* P \tag{3}$$

The Super Winner is identified as the node with the highest value in π^*.

If G is strongly connected, the random walk converges to a unique stationary distribution π^* regardless of $\pi^{(0)}$. Otherwise, the convergence depends on the initial distribution and the graph's connectivity structure.

3.1 Example

Consider a graph G with four nodes A, B, C, D and transition probability matrix

$$P = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 \\ 1/3 & 0 & 1/3 & 1/3 \\ 0 & 1/2 & 0 & 1/2 \\ 1/2 & 0 & 1/2 & 0 \end{pmatrix} \tag{4}$$

We can verify that the stationary distribution is

$$\pi^* = \begin{pmatrix} 1/6 & 1/3 & 1/3 & 1/6 \end{pmatrix} \tag{5}$$

since $\pi^* = \pi^* P$. Nodes B and C have the highest stationary probabilities of 1/3, so they are identified as the Super Winners.

The convergence rate to π^* depends on the eigenvalues of P. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of P, with $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$. The convergence is faster if $|\lambda_2| \ll 1$.

4 Algorithm

The algorithm for identifying the Super Winner can be summarized as follows:

1. Read the network data from a dataset.

2. Create a directed graph $G = (V, E)$ where each node $v \in V$ represents a student and each edge $(u, v) \in E$ represents a relationship from student u to student v.

3. Perform a random walk on the graph G for a fixed number of iterations.

4. At each step of the random walk:
   - If the current node v has outgoing edges, move to a randomly chosen neighboring node u with probability $p_{vu} = \frac{1}{d_v}$.
   - If the current node v has no outgoing edges, move to a random node in V with equal probability $\frac{1}{|V|}$.
5. Record the number of times each node is visited during the random walk.
6. Identify the node with the highest visit frequency as the Super Winner.

5 Results

After applying the algorithm to the given network data, we identify the Super Winner as student 2023CSB1091 (AADIT MAHAJAN), who has an in-degree of 44. This result suggests that AADIT MAHAJAN is the most influential student within the network, as the random walker visits their node more frequently than any other node during the simulation.

6 Conclusion

The random walk algorithm provides an effective method for identifying influential nodes in a network, particularly in the context of social network analysis. By simulating the movement of a walker through the network and analyzing the visit frequencies of nodes, we can determine the most influential individuals, or Super Winners, within the network.

The approach presented in this report has several potential applications, including targeted marketing, community engagement initiatives, and recommendation systems. By identifying the Super Winners, organizations or institutions can leverage their influence to effectively disseminate information, promote ideas, or drive behavior change within the network.

7 Visualization

Figure 1 shows the visualization of the student network.

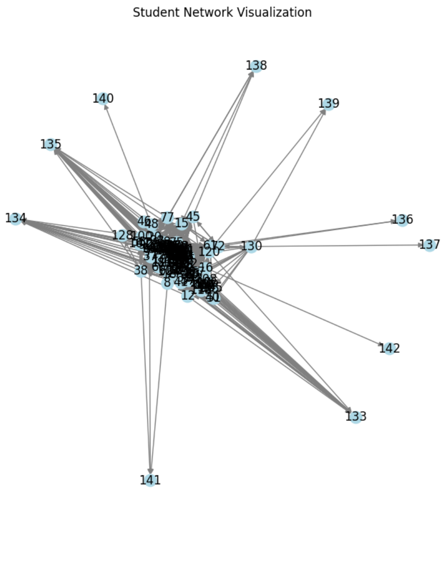

Figure 1: Student Network Visualization

8 Code

The code used to implement the algorithm and generate the results can be found at this link.

References

[1] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[2] Jon Kleinberg. Authoritative sources in a hyperlinked environment. In *ACM-SIAM symposium on Discrete algorithms*, pages 668–677. Society for Industrial and Applied Mathematics, 1999.

[3] David Liben-Nowell and Jon Kleinberg. The four dimensions of social network analysis: An overview of research methods, applications, and software tools. *Bulletin of the American Society for Information Science and Technology*, 33(5):19–23, 2007.

[4] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. *Journal of graph algorithms and applications*, 10(2):191–218, 2006.